Algorithm A.11 Conjugate gradient ascent

Procedure Conjugate-Gradient-Ascent (
  $\boldsymbol{\theta}^1$, // Initial starting point
  f_{obj}, // Function to be optimized
  δ // Convergence threshold
)
  $t \leftarrow 1$
  $\boldsymbol{g}^0 \leftarrow \mathbf{1}$
  $\boldsymbol{h}^0 \leftarrow \mathbf{0}$
  do
    $\boldsymbol{g}^t \leftarrow \nabla f_{\text{obj}}(\boldsymbol{\theta}^t)$
    $\gamma^t \leftarrow \frac{(\boldsymbol{g}^t - \boldsymbol{g}^{t-1})^T \boldsymbol{g}^t}{(\boldsymbol{g}^{t-1})^T \boldsymbol{g}^{t-1}}$
    $\boldsymbol{h}^t \leftarrow \boldsymbol{g}^t + \gamma^t \boldsymbol{h}^{t-1}$
    Choose η^t by line search along the line $\boldsymbol{\theta}_t + \eta \boldsymbol{h}^t$
    $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t + \eta^t \boldsymbol{h}^t$
    $t \leftarrow t + 1$
  while $\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\| > \delta$
  return $(\boldsymbol{\theta}^t)$

A.5.3 Constrained Optimization

In appendix A.5.1, we considered the problem of optimizing a continuous function over its entire domain (see also appendix A.5.2). In many cases, however, we have certain constraints that the desired solution must satisfy. Thus, we have to optimize the function within a constrained space. We now review some basic methods that address this problem of *constrained optimization.*

constrained optimization

Example A.5

Suppose we want to find the maximum entropy distribution over a variable X, with $Val(X) = \{x^1, \ldots, x^K\}$. Consider the entropy of X:

$$\boldsymbol{H}(X) = -\sum_{k=1}^{K} P(x^k) \log P(x^k).$$

We can maximize this function using the gradient method by treating each $P(x^k)$ as a separate parameter θ_k. We compute the gradient of $\boldsymbol{H}_P(X)$ with respect to each of these parameters:

$$\frac{\partial}{\partial \theta_k} \boldsymbol{H}(X) = -\log(\theta_k) - 1.$$

Setting this partial derivative to 0, we get that $\log(\theta_k) = -1$, and thus $\theta_k = 1/2$. This solution seems fine until we realize that the numbers do not sum up to 1, and hence our solution does not *define a probability distribution!*

The flaw in our analysis is that we want to maximize the entropy subject to a constraint *on the parameters, namely, $\sum_k \theta_k = 1$. In addition, we also remember that we need to require that $\theta_k \geq 0$. In this case we see that the gradient drives the solution away from from 0 ($-\log(\theta_k) \to \infty$ as $\theta_k \to 0$), and thus we do not need to enforce this constraint actively.* ■